DOCUMENT RESUME

ED 185 059 TM 800 037

AUTHOR McLarty, Joyce R.
TITLE Multi-Level Item Analysis.
PUB DATE 11 May 79
NOTE 40p.; Paper presented at the Annual Meeting of the California Society of Educational Program Auditors (5th, San Francisco, CA, May 11, 1979). Figure 1 is marginally legible due to small print size.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Variance; Biology; Correlation; Group Membership; Individual Differences; *Item Analysis; Junior High Schools; *Sampling; Science Tests; Test Interpretation; *Test Items
IDENTIFIERS International Assn Evaluation Educ Achievement

ABSTRACT

A new approach to item analysis, concerned not only with differences among individuals, but also differences among the groups in which the individuals may have membership, is described. This multi-level aspect may be applied to studies of differences among individuals, classes, and schools, depending upon the kind of problem being studied and the manner in which the sample is obtained. Most educational data are multi-level: data are collected from individuals (students, teachers) within groups (classrooms, schools). Observed group effects may be the result of the ways in which individuals are selected into the groups or of the effects on individuals of the programs and processes which take place within the groups. Empirical item analysis usually combines information from students in multiple groups, resulting in reliability and validity coefficients which combine between and within group relationships. Application of multi-level regression techniques to item analysis is presented, using biology test data for the United States from the International Association for the Evaluation of Educational Achievement Six-Subject Survey. Statistical procedures for disaggregating the between and within group relationships are presented, and possible interpretation of the results and implications for program evaluation are discussed. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED185059

Multi-Level Item Analysis

Joyce R. McLarty

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. McLarty

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Fifth Annual Conference of the California Society of Educational Program Auditors and Evaluators, San Francisco, May 11, 1979.

TM 800037

Abstract

Most educational data are multi-level; data are collected from individuals (students, teachers, etc.) within groups (classrooms, schools). Observed group effects may be the result of the ways in which individuals are selected into the groups or of the effects on individuals of the programs and processes which take place within the groups. Empirical item analysis usually combines information from students in multiple groups, resulting in reliability and validity coefficients which combine between and within group relationships. This paper presents an application of multi-level regression techniques to item analysis using Biology test data for the U.S.A. from the IEA Six-Subject Survey. Statistical procedures for disaggregating the between and within group relationships are presented, and possible interpretations of the results and implications for program evaluation are discussed.

Most educational data are multi-level; schools are aggregates of classrooms, classrooms of the persons and processes within them. Burstein points out that typically there are substantial differences in the relationships among variables across different levels of aggregation for the same analytic model. In general, he observes, "aggregation ... inflates the estimated effects of background on outcomes and decreases the liklihood of identifying teacher/classroom/school characteristics and practices that are effective." (Note 1, p.86). The results of evaluation and research on classrooms and schools depend in part on the level of aggregation, the unit of analysis, which is employed in the analysis.

Since analytic models using test-score variables vary with the level of aggregation, it is reasonable to suspect that the models may also differ for the individual items from which the test scores are derived. Cronbach, in a parenthetical but penetrating aside in his 1976 monograph titled Research on Classrooms and Schools, appears to have been the first to recognize the possibility:

> Once the question of units is raised, all empirical test construction and item-analysis procedures need to be reconsidered. Is it better to retain items that correlate across classes? or items that correlate within classes? A correlation based on deviation scores within classes indicates whether students who comprehend one point better than most students also comprehend the second point better than most - instruction

> being held constant. A correlation between classes indicates whether a class that learned one thing learned another, but this depends first and foremost on what teachers assigned and emphasized. This leads us to regard the between-group and within-group correlations of items as conveying different information, and makes the overall correlation for classes pooled an uninterpretable blend. (p. 19.19 - 19.20)

This paper attempts to follow Cronbach's lead by applyiing regression-based approaches to multi-level analysis to the problems of item analysis. In effect, this method decomposes the traditional correlation coefficients which are used to estimate internal-consistency reliability and criterion-related validity at the item level into their between and within group components. In a demonstration of this technique, nine Biology items are analyzed using data on 1210 students attending 50 schools. The paper concludes by identifying problems and potential uses associated with multi-level item analysis and suggesting areas in which further research might prove beneficial.

Item Analysis

In traditional approaches to empirical item analysis, student responses to items are examined to determine whether the items may be flawed in some way, and whether they appear to be related to the test as a whole and to the trait being assessed. A common way of looking at this data is to compare the responses of students who did well and poorly on the test or on some criterion measure of interest. A greater proportion of students who did well should have

responded correctly to each item. Correlation coefficients between the test or criterion score and the item score in question are often used to quantify the relationship. These have the advantage of treating the test or criterion score as a continuous variable, rather than having to dichotomize it in order to categorize students as having done well or poorly.

The relationship between the total test score and the score on a single item from that test is sometimes termed the item's internal-consistency reliability coefficient. The average of item reliabilities provides an estimate of the internal-consistency reliability of the test. The relationship between a score on some measure other than the test of which the item was a part and the item score is termed the item's criterion-related validity coefficient. An average of item validities with respect to the same criterion measure provides an estimate of the criterion-related validity of the test for that criterion. The statistical treatments are the same except that for reliability estimates it is necessary to in effect compute the total test score as it would have appeared without that particular item.

Sirotnik (Note 2) distinguishes between two analytic phases of the research process: the psychometric phase and the study phase. The psychometric phase refers to analyses aimed at assessing and developing the measurement devices. The study phase refers to "efforts which take the measures as given and pursue statistical analyses relevant to the research questions. " (p. 1). Item reliability and validity studies are usually performed in conjunction with the psychometric

phase. Preliminary data on student responses is analyzed and items are revised or selected on the basis of the data and other considerations. Usually this formal a test development process is used for tests which are intended for multiple administrations: in several classrooms, schools, or programs, or on several occasions. Appropriately, pilot test data is collected from a sample of these groups. Analysis of this data, however, typically ignores the existence of the grouping factors. It is only later, in the study phase, when the qualities of the measurement devices are taken as given, that consideration of grouping factors is likely to be incorporated. If there are important differences in the relationships among the test items and between them and the criterion measures depending on the level of aggregation of the data, then the qualities of the measurement device which were established in the psychometric phase may be different than those which actually obtain in a study phase analysis at some other level of aggregation. A test which has been determined to be reliable and valid based on data from individuals without consideration of group membership may not be either reliable or valid when it is used to assess relationships between or within these groups. For this reason, it is important to find out whether item analyses at different levels of aggregation do indeed yield different results and, if so how these results might be used in the psychometric phase to improve measurement devices and in the study phase to improve the accuracy of interpretation of research and evaluation results.

Data Aggregation

Item analysis which considers data at several levels of aggregation may be termed "multi-level" in the tradition of multi-level research. "Multi-level" refers to the various levels of data aggregation: data on a classroom is an aggregate of data on individual students, data on a school is an aggregate of data on individual classrooms, and so on. Aggregate data is formed as an average across data for the individual observations. Levels of aggregation usually correspond to conceptual or physical groupings of individuals.

When data from individuals in one or more groups is analyzed irrespective of group membership, the analysis is termed "Total". If, on the other hand, the analysis is based on the group averages, ignoring the individual observations of which they are composed, the analysis is termed "Between groups" or simply "Between". A third type of analysis, called "Within groups" or simply "Within", is based on differences of individual scores from their own group means. In effect, this type of analysis holds the group effects constant while looking at individual difference information. As it turns out, the Total analysis results are a weighted combination of the Between and Within groups relationships. This follows directly from the partitioning of the total sum of squared deviations from the grand mean into its statistically independent between and within groups sums of squares.

Robinson (1950) derived the functional relationship among the total, between, and within group correlations and the magnitude of mean differences between the groups. This

formula can be written as follows (Knapp, 1977):

$$\rho_t = \eta_x \eta_y \rho_b + \sqrt{1-\eta_x^2}\sqrt{1-\eta_y^2}\,\rho_\omega$$

where

ρ_t = the total correlation across N pairs of individual observations

ρ_b = the between correlation computed across G pairs of group means, each weighted by its own n_g

ρ_ω = the within correlation computed across N pairs of deviation scores (each expressed as a deviation from its own mean)

η_x^2 = the correlation ratio, eta^2, between groups for variable X

η_y^2 = the correlation ratio, eta^2, between groups for variable Y

Both eta^{2}s are computed as the ratio of between group to total variation and thus serve as indices of the relative magnitude of mean differences among the groups. As eta^2 for either X or Y approaches zero, the total correlation approaches that within groups. As the eta^2 approaches 1, the total correlation approaches that between groups. In practice this means that where there are small differences between groups, for example where the groups have been matched in some way, the total correlation will approach the correlation within the groups. Where there are large differences between groups, for example where very different intact groups have been compared, the total correlation will approach that between the groups. "In the more interesting and more typical cases of moderate values of between and within variation, all sorts of possibilities

occur." (Sirotnik, Note 2, p. 5). Holding the within group eta^2 for both X and Y variables constant, Sirotnik demonstrated that substantial between group correlations can be obtained under conditions of both homogeneity and heterogeneity among groups, and that the same is true for zero between group correlations. Clearly, the different levels of aggregation convey different information.

Regression Approach

Duncan, Cuzzort, and Duncan (1961) extended the decomposition of effects into their between and within group components to the regression format. Their formula is:

$$B_t = B_b \eta_x^2 + B_w (1-\eta_x^2)$$

where B_t , B_b , and B_w are the total, between group, and pooled within group regression coefficients, and eta^2 is the correlation ratio of X. The regression coefficients can be related to the correlation coefficients and standard deviations as:

$$B_t = \frac{C_{xy}}{S_x^2} = r_{xy}\frac{S_y}{S_x} \qquad B_b = \frac{C_{\bar{x}\bar{y}}}{S_{\bar{x}}^2} = r_{\bar{x}\bar{y}}\frac{S_{\bar{y}}}{S_{\bar{x}}} \qquad B_w = \frac{C_{x_w y_w}}{S_{x_w}^2} = r_{x_w y_w}\frac{S_{y_w}}{S_{x_w}}$$

where 'c' stands for the covariance, 's^2' for the variance, and 's' for the standard deviation, 'r' is the correlation coefficient, the subscript 'w' means "within" and a bar over the letter (as $\bar{X}$) indicates the group mean (for X). The Duncan, Cuzzort and Duncan formula is particularly useful because it indicates clearly the relative weights of the between and within group components in the total relationship. Use of an unstandardized regression coefficient makes the

comparison of relationships for the between and within group components within an item and across items for a single criterion somewhat easier. Use of standardized regression coefficients (correlation coefficients in this case) is necessary for comparison of relationships across multiple criterion measures for the same or different items.

The technique for estimating the between and within groups components of item scores in a regression format involves generating two variables from the single item score variable and the group membership information. The between group variable consists of the group means on each item, weighted by the number of observations in the group. This, in more familiar terms, is the difficulty level of the item for the group. The weighting maintains the correspondence of the analysis to the formulas given by Robinson and by Duncan, Cuzzort, and Duncan.

The within group variable is computed by subtracting the relevant group mean from each individual's score on an item by item basis. The group mean score for an item is a constant for the group and, in this case, the item score is binary (zero or one), so these deviation scores also turn out to be binary (zero or one minus the group mean). Because the group mean is statistically independent of deviations from it, the regression analysis for the between and within group analyses may be performed simultaneously. The criterion score, Y, is regressed on the weighted group mean and the individual's deviation from it. If the analysis is based on the correlation matrix of the variables (standardized)

the resulting regression coefficients are interpretable as correlation coefficients. If the regression is in unstandardized form, the resulting relationships are in the scale of the Y variable. In either case, the statistical significance of the coefficients (or lack of it) will be the same.

Data Analysis

In order to explore the problems and potential of the regression approach to multi-level item analysis, a data set containing information on individual students' responses to multiple choice items was selected for analysis. The data used was a subset of items of the International Association for the Evaluation of Educational Achievement (IEA) Six-Subject Survey. Science data were selected as potentially more indicative of school-based learning than the other available subjects. Details of the Science Survey and of the method by which the Science Test items were constructed and selected is available in the report by Comber & Keeves (1973). Items were generated by the countries involved, pilot-tested in a sample of the countries, and selected on the basis of traditional characteristics (internal-consistency reliability, properly functioning distractors, etc.). Although the level of analysis used in selecting the items is not reported, some within country statistics were apparently used.

The Biology subtest data for Population II (14 year olds) in the United States was used. Data on the results of the IEA study in the USA are available in the report by Wolf (1977). It was expected that Biology would show more variation in instruction at this age level than would Chemistry or Physics. Since two forms

of the Science Test had been administered, Form B was selected at random for analysis. It contained 9 Biology items. These items are reproduced as Figure 1. Their original numbering (items 2 through 10) is retained in this report for ease of cross-reference.

The original data set contained information for more than 3,500 students nested in nearly 200 schools. For economic reasons, a sample was drawn from this set. Since accurate estimation of group means was central to the analysis, and since schools were the only appropriate grouping factor available (no classroom grouping information was available), sampling was on the basis of schools rather than students, and schools with fewer than 20 students having Biology Form B data were not selected. With these restrictions a sample of 50 schools (of the eligible 110) was selected at random. The final data set included information on for 1210 students, all of whom were 14 years old and most of whom (71%) were in the 9th grade. Forty-nine percent were boys, fifty-one percent girls. No ethnic data were available, and socio-economic data were not used.

Each of the biology items was scored according to the key provided. Items with no response were scored as incorrect since no student in the sample had failed to take the test and since the placement of the biology items (#2-10) on the 40 item Science Test assured that the students would have reached them even if they had not completed the test. Fewer than 2% of the students had failed to respond to any given item. The school mean was calculated for each school on each item and added to the information for each student in that school.[1] That mean was then subtracted from the student's

score on an item by item basis. The deviation scores corresponding to the items were also added to the information for each student.

Biology test scores were computed for use in the reliability analysis by constructing a separate score for each item. That is, the "Total Biology test score" for item #2 was the sum of the items answered correctly for items #3-10. For item #3, it was the sum of the correct items for items #2 and #4-10, and so on. This gave 9 separate test scores, one paired to each item, providing criterion variables with the influence of the item under consideration removed.

For the validity analyses, a number of available criterion measures were selected. Three of them are reported here:

1. Opportunity. Each school had provided information on the proportion of students at that school who had had the opportunity to learn the information covered on each item. After reflection of the scale for ease of interpretation, the five-point scale was:

   1 - no one had the opportunity
   2 - fewer than 25% had the opportunity
   3 - 25% - 75% had the opportunity
   4 - more than 75% had the opportunity
   5 - everyone had the opportunity

   Opportunity data was paired to the items. Since the data were at the school level, there was no within school variation. Data was missing for 12 of the 50 schools.

2. Raw Word Knowledge (RWK). This was a raw score (number of items correct) on a 40-item vocabulary test. It probably indicates language ability most specifically, and general ability as well. It may also be related to

such non-school factors as socio-economic level and race. It may also reflect the general level of the school's English program.

3. Liking of Biology (LIKEBIO). Students were asked to rate each school subject on a five-point scale. Five was "a favorite", one "least liked". No comparison of subjects was involved, each was rated separately.

Although the data for the test items were complete for the 1210 students, data on some of the other variables were missing in some cases. Analyses were computed on the basis of the cases having complete information for the variables included. This accounts for the varying Ns. Data for both standardized (correlation-based) and unstandardized regressions are presented. Because of the large N's, any standardized regression coefficient over about .05 is statistically significant. The corresponding unstandardized coefficient is, of course, significant as well. The R^2 for each analysis can be calculated by squaring the standardized coefficient reported.

Results - Descriptive

Item means for the total group ranged from .19 for item 5 to .71 for item 4. Standard deviations were much greater within groups than between, and this is reflected in the relatively low eta^2 ratios (.06 to .10). Because of the low proportion of variance associated with group membership in this case, the total regression coefficient can be expected to mimic the within group relationship. (see Table 1).

Item intercorrelations (Table 2) were very low at the total and within group levels of analysis. So low as to suggest that the items were not measuring a single construct. The low within group correlation means that students who had learned the answer to one biology item did not necessarily know the answer to another one. This may be indicative of the effect of using a small number of items to estimate a large construct ("Biology knowledge"). The internal-consistency reliability of this set of items for estimating within school differences in biology knowledge is very low.

Between schools the inter-item correlations are somewhat higher, but still not impressive. What relationship there is among the items appears to stem from the fact that when students at one school respond correctly to an item in general they will respond correctly to others. These items thus provide a reasonably reliable indicator of the general level of biology knowledge for a school. This information is extended in the item by item reliability analysis (Table 3). While the item intercorrelations look at pairs of items, the reliability analysis looks at the relationship between each item and the remaining set. In general, the regression coefficients are larger between groups than within, a fact which is obscured by the total relationship. For example, for item #3, the unstandardized regression coefficient between groups is 2.87 compared to only .56 within. Weighted as they are in the total regression by eta^2 and $(1-eta^2)$, however, the between group effect contributes about .30

(2.87 x .1041) while the within group effect contributes about .50 (.56 x .8959) to the total coefficient (differences are due to rounding errors).

Table 4 illustrates what can happen when data analysis is conducted at the wrong level. The total regression coefficients, those which would have been calculated had grouping information not been included, are very low. One would have concluded either that there was very little relationship between the opportunity to learn and the actual learning having taken place, or that the schools had provided very poor opportunity information. Examination of the between group information suggests that there is, for most items, a reasonable relationship between the opportunity to learn and the score. Given the school-wide nature of the estimate, and the crudity of the scale, these relationships are relatively strong. Strong enough for the negative information for item #8 to be puzzling. For that item, it appears that the greater the opportunity for the students to have learned the answer, the less likely they were to have given it! Unfortunately the curriculum information which might explain this situation is not available.

The Raw Word Knowledge analysis, presented in Table 5, again shows relatively stronger item-criterion relationships between than within groups. Comparison of items #3 and #8, which would have appeared to have the same degree of relationship to the RWK score had only the total analysis been considered, indicates that appearances can be misleading. Item #3 has a far stronger relationship between groups to RWK than it does within groups, while item #8 has similar relationships

to RWK at both levels. In words, this suggests that the student attending a school where a large proportion of students responded correctly to item #3 is likely to have a fairly high word knowledge score. If, however, the student responded correctly to item #3 when many of the students at that school did not (that is, the student had a large deviation score), this is not as predictive of a high vocabularly level. The level of the school on this item is more predictive of a high level of word knowledge than is individual ability. For item #8, both are equally predictive.

A similar comparison can be made for the criterion LIKEBIO. For item #3, more of the relationship is between groups than within. If one attends a school with a relatively high mean on item #3, one is more likely to indicate a liking for biology. Having responded correctly to item #3 when students at the school in general did not is somewhat less predictive of this attitude. For item #8, however, the reverse is true. Attending a school where a large proportion of students responded correctly to item #8 does not necessarily predict that one likes biology. However having responded correctly to item #8 when most of one's cohort did not is associated with liking biology.

Results - Interpretive

Making sense out of the empirical analysis in the absense of information about the curriculum is difficult. One can at best theorize about the meanings that might be associated

with these patterns of numbers. For simplicity's sake, only two items will be discussed here: item #3 and item #8 (Table 7). They were chosen because they appear to have somewhat different patterns of relationships to the other measures, despite similar variances between and within groups and similar correlation ratios. Item #3, which was classified by IEA as measuring Applications of science, asks the student to identify a feature of an animal skull which might lead one to conclude that the animal "preyed on other animals for its food". It was relatively easy, with a mean score for the total group of .65. Item #8, classified as testing Information, was much more difficult with a mean score of .25. It asked for the name of the substance which an active muscle, "that is, a muscle which is doing work" gives up to the blood. Examination of the items indicates that item #3 is much wordier than item #8, and that it is more amenable to being "figured out" as opposed to being remembered. A student might theorize that an animal would require "long and pointed teeth" to prey on other animals, or might visualize and classify familiar animals which prey on others and thereby arrive at the correct answer by elimination of the distractors. Neither approach seems particularly likely to result in a correct response for item #8 where one needs to know that carbon dioxide is given up to the blood.

Item #3 is uniformly more sensitive at the between school level than at the within school level. Attending a school where many students responded correctly to item #3 was related to a relatively high level of vocabulary, relatively greater

opportunity to study biology content, and a relatively greater liking of biology. Having done well on item #3 when the students at one's school did relatively poorly was also indicative of these things, but less so. It was the level of the school as a whole which was most predictive.

For item #8, attending a school where the students did relatively well on the item was no more indicative of a high vocabulary level than having done well when they did not. In fact, having done well on item #8 when the school mean on that item was low (having been one of the few students who responded correctly) was related to liking biology more strongly than attending a school where many of the students responded correctly to that item. Attending such a school was related to having had, on the average, less opportunity to study biology. Such a situation might have occurred because the school, while offering biology instruction to only a few students, offered those students a program which included information on "substances given off by working muscles". Because of the format of the opportunity information, a small percentage of students offered this information might be associated with a relatively high school mean (since at other schools more students might have had the opportunity to learn this information but in a less demanding program so that they did not, in fact, learn it).

Such conjecture in the absense of curricular information can be neither confirmed not denied. If the information in this analysis generalizes to other data sets and other items (and there is no guarantee that it will), one might expect

a fact-oriented test item to relate to other measures in terms of not only the general level of the group (program effects), but individual differences from that level as well (ability factors). Application-oriented items, on the other hand, would show more sensitivity between groups (program effects, or perhaps indices of the general level of the community) and relatively less relationship based on individual differences within the group. These are only hypotheses born of a single exploratory study using less than ideal data and less than ideal groupings of students. Much more work remains to be done.

Directions for Further Efforts

Although the statistical techniques used in regression-based multi-level item analysis are relatively simple, some of the technical considerations are not. Questions which need to be investigated include: when should within-group information be pooled and when not? How should the variables be scaled for analysis? Does classical test theory ($X = T + E$) hold up when data is aggregated? What analysis can be done when more than one level is nested (students within classrooms within schools) or when grouping variables are crossed (boys and girls within classrooms)?

Conceptual questions are even more puzzling. Perhaps the most important is: having conducted item analysis between and within groups and having recognized that the results are different, which should one use? At present the answer seems to be that the decision must be made on conceptual grounds. If that which is to be measured is considered to

be a global effect (classroom climate, program effectiveness) which is measured through individuals only because that is the only available route, then the between group relationships are probably the most relevant. Individual differences are inaccuracies in perception (or reception of the effect) and should "wash out" in the analysis.

If, on the other hand, the effect on individuals within the group is what is being measured (improved self-image, effectiveness of individualized instruction) then the within-group approach may be preferable. The conceptual distinction is not clear-cut, and more thought is needed.

Summary of Results

What does appear to be clear-cut, if anything is, is that between and within group analyses give different results. Use of the total analysis is appropriate only when all grouping factors are considered irrelevant. Because of this, qualities of measurement devices estimated from the total analysis will be inaccurate estimates of their ability to give reliable and valid information between or within groups. If the eta^2 for items is, in general, about .06 to .10 as it is for this set, selection of items on the basis of the total analysis will result in tests which are maximally reliable and valid with respect to individual differences within groups and minimally so for differences between groups. This would be expected to lessen the ability to detect between group differences while improving the estimation of differences within groups (those based on individual ability or interest).

Recommendations

Every research paper seems to end with a recommendation for further research. While this is especially appropriate in a technically new-born area such as this, and while some directions in which additional efforts are needed have been suggested above, further research seems perhaps too bland a recommendation for such an important topic. What can evaluators do to improve their measurements devices and to make more accurate interpretations of their results right now? First, use some common sense. If you are constructing a test to measure differences between groups, and if you have enough data to make it possible, do a between-groups item analysis, and use the results as part of your item revision and selection process. If you are working with an existing instrument, calculate the eta^2 for your groups. If it is small, the analysis you have done is probably close to the within group relationship. Is that what you want? If not, you may be missing something important.

Do you expect your programs to have important effects on the variability of student performance within the classroom? Perhaps aptitude by treatment interactions are expected? Maybe the program is designed to be compensatory, or for individualized instruction (decreasing or increasing variation within the class)? You need the within group information. If your eta^2 is large, you are probably not getting it with a total analysis.

Are you trying to construct instruments which are less sensitive to ethnic or language or sex characteristics? If the instructional program is held constant (students are in the same

instructional program) you can look for items which derive their reliability and validity from relationships within groups rather than between them. If, on the other hand, the purpose is to construct an instrument which is maximally reactive to instruction, do the grouping on the basis of an exposure measure, and look for items with high reliability and validity between the groups rather within them. The characteristics of items with the desired statistical patterns may give you clues as to the differences between and within groups.

Consideration of between and within group relationships is appropriate for all of your quantatative data, and for most statistical analyses. Use multi-level approaches for constructing attitude and rating scales, for analyzing patterns of subjective grading, for understanding research results. Use it in regression analysis, correlational studies, and factor analysis. Whatever data collection and analysis you do, if grouping is relevant (and it usually is) think about whether what you are doing is appropriate to your purposes.

If what we have learned so far is correct, doing analysis without consideration of grouping factors makes us less able to construct instruments which are sensitive to effective educational programs. As program evaluators, it is especially important for you to be aware of this. Whether you are constructing instruments or interpreting their results, do it in the way which is most likely to give you the information you require. Chose the most appropriate level for your analysis.

Reference Notes

1. Burstein, L. The role of levels of analysis in the specification of educational effects. Draft of a paper commissioned by the Educational Finance and Productivity Center, Department of Education, University of Chicago, under contract from the National Institute of Education, June, 1978.
2. Sirotnik, K. Psychometric implications of the unit-of-analysis "problem" (with examples from the measurement of organizational climate). Draft of a paper produced in association with the Institute for Development of Educational Activities, March, 1979.

References

Comber, L. C. & Keever, J. P. Science education in nineteen countries: international studies in evaluation I. Stockholm: Almqvist & Wiksell, 1973.

Cronbach, L. J. Research on classrooms and schools: formulation of questions, design, and analysis. Stanford: Stanford Evaluation Consortium, 1976.

Duncan, O. D., Cuzzort, R. P. & Duncan, B. D. Statistical geography. Glencoe: Free Press, 1961.

Knapp, T. R. The unit-of-analysis problem in applications of simple correlation analysis to educational research. Journal of Educational Statistics, 1977, 2, 171-186.

Robinson, W. S. Ecological correlations and the behavior of individuals. American Sociological Review, 1950, 15, 351-357.

Wolf, R. M. Achievement in America: national report of the International Educational Achievement Project. New York: Teachers College Press, Columbia University, 1977.

Footnotes

1. This and all other calculations were performed on the U.C.L.A. computer using the SPSS subroutine package and data manipulation capabilities. Data aggregation proved quite expensive since 450 school means (one per item per school) had to be calculated. The addition of the 10,890 deviation scores to the data set (one per student per item) made the total data set too large for available disk space and it had to be transferred to tape. The regression runs were relatively inexpensive. Regression sets with separate criterion measures were run separately to avoid the elimination of cases unnecessarily (list-wise deletion was used). The use of a composite aggregate file provided the necessary weighting of group means.

Table 1:

Item Descriptive Data
(N=1210)

Item #	Mean	Standard Deviation			eta^2
		Total	Between	Within	
2	.33	.47	.12	.45	.0664
3	.65	.48	.15	.45	.1041
4	.71	.46	.13	.44	.0803
5	.19	.39	.11	.38	.0773
6	.63	.48	.13	.47	.0711
7	.49	.48	.15	.46	.0961
8	.25	.43	.14	.41	.1044
9	.32	.47	.14	.44	.0919
10	.34	.45	.11	.43	.0614

Table 2:

Item Intercorrelations
(N=1210)

Item #	2	3	4	5	6	7	8	9	10
2	(total)								
	(between)								
	(within)								
3	-.02								
	.03								
	-.03								
4	.05	.16							
	.19	.66							
	.04	.11							
5	.05	.06	.02						
	.26	.26	.19						
	.04	.04	.01						
6	.04	.15	.11	.08					
	-.04	.53	.47	.40					
	.04	.12	.09	.05					
7	.04	.16	.20	.03	.25				
	-.01	.58	.52	.29	.53				
	.05	.12	.17	.00	.23				
8	.10	.14	.11	.08	.15	.14			
	-.36	.33	.40	.19	.25	.36			
	.07	.12	.08	.07	.15	.11			
9	.13	.12	.10	.07	.07	.11	.18		
	.43	.49	.44	.29	.37	.41	.38		
	.10	.08	.07	.04	.04	.08	.16		
10	.05	.04	.03	.06	.09	.08	.10	.03	
	.15	.40	.32	.52	.32	.28	.23	.16	
	.05	.01	.01	.03	.08	.06	.09	.02	

Table 3:

Internal-Consistency Reliability Study Using Paired Subtest Scores*

Item #	Subtest		Standardized Betas			Unstandardized Regression				
	Mean	Standard Deviation	Total	Between	Within	Total	Const.	Between	Within	Const
2	3.67	1.70	.12	.11	.09	.42	3.53	1.44	.34	3.19
3	3.35	1.65	.23	.26	.15	.78	2.83	2.87	.56	1.46
4	3.29	1.66	.22	.26	.15	.80	2.73	3.48	.58	.83
5	3.81	1.73	.12	.18	.07	.53	3.70	2.77	.33	3.28
6	3.37	1.63	.27	.23	.21	.90	2.79	3.00	.75	1.47
7	3.36	1.62	.29	.24	.22	.97	2.75	2.57	.79	1.74
8	3.74	1.65	.28	.20	.23	1.06	3.48	2.39	.91	3.14
9	3.68	1.70	.22	.24	.16	.79	3.43	2.77	.59	2.80
10	3.72	1.70	.13	.18	.09	.50	3.58	2.84	.35	2.93

*N=1210

70 3

Table 4:

Criterion-Related Validity Study Using Paired Opportunity Score

Item #	Opportunity		Standardized Betas			Unstandardized Regression				
	Mean	Standard Deviation	Total	Between	Within*	Total	Const.	Between	Within*	Const.
2	3.74	1.38	,02	.06	-	.05	3.73	.78	-	3.50
3	3.37	1.47	.06	.22	-	.19	3.24	2.49	-	1.72
4	3.74	1.43	.02	.09	-	.06	3.70	1.34	-	2.78
5	3.42	1.51	.12	.51	-	.45	3.33	8.31	-	1.81
6	3.94	1.37	.10	.39	-	.30	3.74	4.31	-	1.13
7	3.67	1.33	.06	.25	-	.17	3.54	2.75	-	1.87
8	3.50	1.38	-.06	-.20	-	-.18	3.55	-2.13	-	4.04
9	3.04	1.32	-.03	-.10	-	-.07	3.06	-1.00	-	3.37
10	3.17	1.39	.06	.28	-	.19	3.12	3.96	-	2.07

*Since data was at the school level, there was no within school variance.

N=830

Table 5:

Criterion-Related Validity Study Using RWK* Score

Item #	Standardized Betas			Unstandardized Regression				
	Total	Between	Within	Total	Const.	Between	Within	Const.
2	.05	.11	.02	.58	27.40	4.82	.28	26.00
3	.24	.27	.17	2.69	25.83	9.59	1.95	21.33
4	.25	.28	.18	2.90	25.53	11.96	2.17	19.12
5	.12	.16	.07	1.58	27.29	7.77	1.05	26.11
6	.17	.23	.11	1.87	26.41	9.78	1.29	21.40
7	.23	.26	.16	2.56	25.97	9.22	1.84	21.75
8	.24	.19	.19	2.92	26.87	7.29	2.43	25.77
9	.17	.26	.09	1.89	27.00	9.87	1.09	24.47
10	.09	.12	.07	1.12	27.28	5.70	.84	26.01

*Raw Word Knowledge: Mean= 27.78, Standard Deviation= 5.21, N=1183

Table 6:

Criterion-Related Validity Study Using LIKEBIO* Score

Item #	Standardized Betas			Unstandardized Regression				
	Total	Between	Within	Total	Const.	Between	Within	Const.
2	.03	.01	.03	.05	3.24	.06	.05	3.23
3	.13	.14	.09	.19	3.12	.66	.14	2.82
4	.09	.10	.06	.14	3.15	.59	.10	2.84
5	.01	.10	-.02	.02	3.25	.65	-.04	3.13
6	.12	.04	.11	.18	3.14	.21	.17	3.12
7	.13	.07	.12	.20	3.12	.35	.19	3.03
8	.18	.08	.16	.30	3.18	.40	.29	3.15
9	.12	.08	.10	.19	3.19	.38	.17	3.13
10	.12	.08	.11	.20	3.20	.53	.18	3.11

*LIKEBIO: Mean= 3.25, Standard Deviation= .73 , N=1166

Table 7:

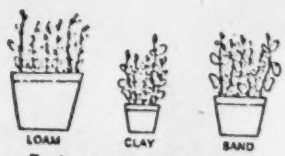
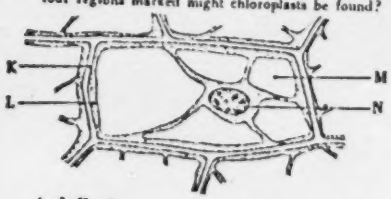
Comparison of Items #3 and #8

Statistic	Item #3	Item #8
Mean (N=1210)	.65	.25
Standard Deviation		
Total	.48	.43
Between	.15	.14
Within	.45	.41
Eta^2	.1041	.1044
I-C Reliability		
Total	.23	.28
Between	.26	.20
Within	.15	.23
R w/Opportunity		
Total	.06	-.06
Between	.22	-.20
R w/ RWK		
Total	.24	.24
Between	.27	.19
Within	.17	.19
R w/Likebio		
Total	.13	.18
Between	.14	.08
Within	.09	.16

* Figure 1:

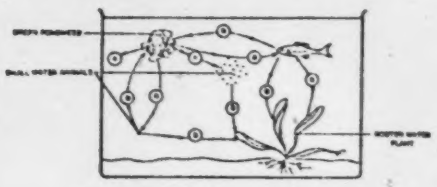
Biology Test Items Form B

Reprinted with permission of the International Association for the Evaluation of Educational Achievement.

	Target Popul.	Content Area	Beha-vior	Average Facility	Med. Discrim.	Popul. Discrim.	Effect. Distr.	Easier in	Harder in
Population II–Test 4 B									
1. The sun is the only body in our solar system that gives off large amounts of light and heat. We see the moon because it is • A. reflecting light from the sun. B. without an atmosphere. C. a star. D. the biggest object in the solar system. E. nearer the earth than the sun.	II (1A.1)	Earth (1)	Under.	67.7	33	–	–	Italy	–
2. In an experiment green leaves were put in a jar and the apparatus was kept in the dark. Lime water was turned cloudy by the gas that formed in the jar. Which of the following gives the best explanation of this result? A. O_2 was produced by photosynthesis B. O_2 was produced by respiration • C. CO_2 was produced by respiration D. O_2 was used up in respiration E. CO_2 was produced by photosynthesis	II	Biol. (13)	Under.	36.4	27	E, D	E	India Iran	FRG Finland Netherlands
3. John brought the skull of an animal to school. His teacher said she did not know what the animal was but she was sure that it was one that preyed on other animals for its food. Which clue, do you think, led her to this conclusion? A. The eye sockets faced sideways B. The skull was much longer than it was wide C. There was a projecting ridge along the top of the skull • D. Four of the teeth were long and pointed E. The jaws could work sideways as well as up and down	II (1B.6)	Biol. (12)	Applic.	63.3	33	–	–	Belgium (Fr) Finland Italy	Scotland Thailand
4. Tom wanted to learn which of three types of soil—clay, sand, or loam—would be best for growing beans. He found three flowerpots, put a different type of soil in each pot, and planted the same number of beans in each, as shown in the drawing. He placed them side by side on the window sill and gave each pot the same amount of water.  Why was Tom's experiment *NOT* a good one for his purpose? A. The plants in one pot got more sunlight than the plants in the other pots • B. The amount of soil in each pot was not the same C. One pot should have been placed in the dark D. Tom should have used different amounts of water E. The plants would get too hot on the window sill	II (1A.8)	Biol. (13)	Under.	54.8	29	–	–	Belgium (Fr) Netherlands USA	Belgium (Fl) FRG Hungary Japan
5. The drawing represents a plant cell. In which of the four regions marked might chloroplasts be found?  A. In K only • B. In L only C. In M only D. In N only E. In both K and L	II	Biol. (14)	Inform.	17.1	17	D	C	Finland	Hungary
6. The energy for photosynthesis is generally obtained from A. chlorophyll. B. chloroplasts. • C. sunlight. D. carbohydrates. E. carbon dioxide.	II (11B.1)	Biol. (13)	Inform.	62.1	25	–	–	Chile India Iran New Zealand Thailand	Belgium (Fl) Belgium (Fr) FRG

BEST COPY AVAILABLE

Biology Test Items - Form B

	Target Popul.	Content Area	Behavior	Average Facility	Med. Discrim.	Popul. Discrim.	Effect. Distr.	Easier in	Harder in
7. The diagram below shows an example of interdependence among aquatic organisms. During the day the organisms either use up or give off ⓐ or Ⓑ as shown by the arrows.  Choose the right answer for ⓐ and Ⓑ from the alternatives given. • A. ⓐ is oxygen and Ⓑ is carbon dioxide B. ⓐ is oxygen and Ⓑ is carbohydrate C. ⓐ is nitrogen and Ⓑ is carbon dioxide D. ⓐ is carbon dioxide and Ⓑ is oxygen E. ⓐ is carbon dioxide and Ⓑ is carbohydrate	II (11B.2)	Biol. (13)	Under.	48.7	36	–	–	Thailand	Netherlands
8. What does an active muscle, that is, a muscle which is doing work, give up to the blood? • A. Carbon dioxide B. Oxygen C. Nitrogen D. Vitamin B E. Glucose	II	Biol. (14)	Inform.	21.8	31	B, D, E	C	Hungary Italy	Japan
9. The Andes are high mountains in South America and their inhabitants live and work at high altitudes. These people have almost twice as many red corpuscles in their blood as do the people living in the valleys. Which one of the following is the best explanation of this? A. In the Andes there is less air pressure acting on the inhabitants' blood vessels and so new red corpuscles can be produced more quickly B. Because there is a smaller amount of oxygen in the air of the Andes the inhabitants breathe more deeply in order to increase the total amount of oxygen in their lungs • C. In the Andes there is less oxygen entering the lungs of the inhabitants so that an increase in the number of red corpuscles enables a larger proportion of this oxygen to be absorbed D. Inhabitants of the Andes need more red corpuscles to transport oxygen through the blood vessels because there is less oxygen in the air they breathe E. The lower air pressure in the Andes causes blood to circulate more quickly through the blood vessels and so more red corpuscles are needed to transport the oxygen	II	Biol. (17)	Higher	20.8	19	D, E	D	India Iran	Belgium (Fl) Sweden
10. All of the following are aspects of the reproductive process. Which one of them must occur before we can be certain that fertilization has taken place? A. A male organism must find a mate B. Reproductive organs must be produced • C. The nucleus of a male gamete must fuse with that of a female gamete D. A spermatozoon must reach an egg cell E. A female gamete must have a store of food for the embryo	II (10A.7)	Biol. (16)	Inform.	33.9	18	D	D	Belgium (Fl) Iran	Japan
11. Paint applied to an iron surface prevents the iron from rusting by A. preventing nitrogen from coming in contact with the iron. B. reacting chemically with the iron. C. preventing carbon dioxide from coming in contact with the iron. D. making the surface of the iron smoother. • E. preventing oxygen and moisture from coming in contact with the iron.	II (1A.12)	Chem. (23)	Inform.	74.4	35	–	–	–	India Japan

BEST COPY AVAILABLE